Graph-based representation for multiview images with complex camera configurations

Xin Su, Thomas Maugey, Christine Guillemot

GRAPH-BASED REPRESENTATION FOR MULTIVIEW IMAGES WITH COMPLEX CAMERA CONFIGURATIONS

Xin SU, Thomas MAUGEY and Christine GUILLEMOT

INRIA, Rennes 35042, France

ABSTRACT

Graph-Based Representation (GBR) has recently been proposed for rectified multiview dataset. The core idea of GBR is to use graphs for describing the color and geometry information of a multiview dataset. The color information is represented by the vertices of the graph while the scene geometry is represented by the edges of the graph. In this paper, we generalize the GBR to multi-view images with complex camera configurations. Compared with previous work, the GBR representation introduced in this paper can handle not only horizontal displacements of the cameras but also forward/backward displacements, rotations etc. In order to have a sparse (i.e., easy to code) graph structure, we further propose to use a distortion metric to select the most meaningful connections. For the graph transmission, each selected connection is then replaced by a disparity-based quantity. The experiments show that the proposed GBR achieves high reconstructing quality with less or comparable coding rate compared with traditional depth-based representations, that directly compress the depth signal without considering the rendering task.

Index Terms— Geometry information, graph-based representation (GBR), complex camera configurations, disparity, distortion model

1. INTRODUCTION

The Multi-view plus depth format allows free viewpoint rendering. However, this format generates very large volumes of data, hence the need for intermediate compact representations or efficient compression schemes. Basically, the representation should capture both color and geometry of the multiview images. The color information is typically described by 2D images. The geometry information can be represented explicitly by depth or disparity [1, 2], implicitly by feature correspondences between images [1, 3] or by light fields [1, 4]. In this paper, we focus on the explicit geometry representation.

In the multi-view plus depth (MVD) [5] format, the geometry is represented by a depth map. The depth map is exploited by image-based rendering techniques to render virtual views at any viewpoint [6]. The MVD format yields very large volumes of data which need to be compressed. An extension of the high efficiency video coding standard, namely 3D-HEVC [7], has been proposed. However the lossy compression of depth data may cause color displacement artifacts around foreground objects in the rendering views due to the smoothing of depth edges. To solve this problem, rate-distortion models (e.g., in [8]) have been proposed to guarantee less depth error around edges. Another approach in which depth edges are losslessly encoded has been proposed in [9].

Disparity, as an alternative of depth, describes the scene geometry by the displacement of the same point between two views. In other words, contrary to depth, the disparity is linked to a given view synthesis task (which is usually the case in multiview video coding framework). Knowing the camera parameters, the disparity in each point can be easily derived from the depth information. In multiview video coding (MVC) [10], the disparity is used for the inter-view prediction. More recently, a graph-based representation (GBR) [11] has been proposed, in which the graph connections are derived from the disparity and provide just *enough* geometry information to synthesize the considered predicted views. However only horizontal translations of the cameras are considered in GBR [11].

In this paper, we extend the GBR idea to deal with multiview images with complex camera configurations. Beyond cameras horizontal translations, the proposed GBR representation can handle more complex camera motions, such as forward/backward translations and rotations. In the former GBR, the connections describe the disparity as follows. Each connection of the graph links one pixel and its (horizontal) neighboring pixel (the gap between the two pixels is the disparity). The extension to complex camera configurations is not straightforward since the disparity becomes two-dimensional (horizontal and vertical displacement). In order to circumvent this complexification, we use the concept of *epipolar segment* to keep the disparity one-dimensional. An *epipolar segment* (as shown in Fig.1) is a line segment consisting of all possible projections of a pixel with varying depth. The edge in the GBR (e.g., the blue link in Fig.1) links a pixel in one view and its true projection point in another view. Instead of one connection for one pixel, we horizontally group neighboring pixels to form a segment and only one connection is assigned to one segment. A distortion metric with respect to the reconstruction quality is used to group the pixels. The resulting graph is thus sparse and less costly to transmit. Since the constructed graph connects pixels across the views, it provides more neighboring relations than traditional image plus depth

This work has been funded by the regional council of the Brittany Region.

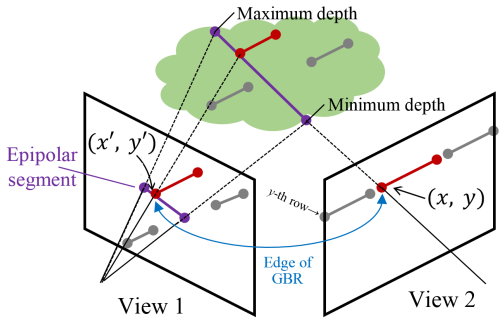


Fig. 1. The concept of GBR: The vertices correspond to pixels of multiview images; The edges link pixels in one view and their projections in another view.

representations, which can be used to better exploit texture redundancy. Thus, although this paper only focuses on the geometry representation, the proposed GBR representation can be considered as a pre-processing step of the future graph-based representation.

2. MULTIVIEW GEOMETRY AND VIEW SYNTHESIS

2.1. Depth Image Based Rendering (DIBR)

Let us consider a scene captured by two cameras $\mathcal{I}_1$ and $\mathcal{I}_2$ of size $X \times Y$ with camera configurations $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$, where $\mathbf{\Phi} = \{\mathbf{M}, \mathbf{R}, \mathbf{T}\}$ are the parameters of the camera. $\mathbf{M}$ is the intrinsic matrix, $\mathbf{R}$ is the rotation matrix and $\mathbf{T}$ is the position of the camera ($[\mathbf{R}| - \mathbf{RT}]$ is also known as the extrinsic matrix). As detailed in [6], pixel (x, y) in view 1 (with associated depth $z_1(x, y)$) can be projected to view 2 by

$$\begin{cases} \begin{bmatrix} x_r \\ y_r \\ z_r \end{bmatrix} = \mathbf{R}_1^{-1}\mathbf{M}_1^{-1} \begin{bmatrix} xz_1(x,y) \\ yz_1(x,y) \\ z_1(x,y) \end{bmatrix} + \mathbf{T}_1 \\ \begin{bmatrix} x'z_2(x',y') \\ y'z_2(x',y') \\ z_2(x',y') \end{bmatrix} = \mathbf{M}_2\mathbf{R}_2 \left(\begin{bmatrix} x_r \\ y_r \\ z_r \end{bmatrix} - \mathbf{T}_2 \right) \end{cases}, \quad (1)$$

where, $[x_r, y_r, z_r]^{\mathrm{T}}$ are the coordinates of the corresponding point in the 3D scene. Under the Lambertian assumption, the color at (x, y) in $\mathcal{I}_1$ is the same as the color at (x', y') in $\mathcal{I}_2$, when (x, y) and (x', y') satisfy Eq.(1).

Considering view 1 as the reference view (the color and depth of view 1 are known), the predicted view (view 2) can be generated by Eq.(1) from the reference view. This is referred to as *forward projection* (or forward warping), which is classically used in depth-image-based rendering (DIBR). On contrary, *backward projection* first computes the depth map of the predicted view. Then the warped depth map is inpainted and used to locate the pixels in the reference view corresponding to each pixel in the predicted view. The color of the reference view is then mapped to the predicted view. In this paper we employ backward projection, since it can easily handle the disocclusions (including *cracks*).

2.2. Depth vs. Disparity

For pixel (x, y) in $\mathcal{I}_1$, its projection (x', y') in $\mathcal{I}_2$ can be located by Eq.(1) with depth $z_1(x, y)$. The disparity of pixel (x, y) is thus defined as

$$\vec{d}(x, y) = (\Delta x, \Delta y) = (x' - x, y' - y) \ , \quad (2)$$

where $\Delta x, \Delta y \in \mathbb{R}$. When the motion of the camera from one view to the other is translational, the geometrical correlation between two views is only horizontal. In this case, the disparity vector $\vec{d}(x, y) = (\Delta x, \Delta y = 0)$ is simplified to a real number $d(x, y) = \Delta x$.

We see here that the fundamental difference between depth and disparity resides in the fact that disparity is associated to two cameras, while depth to only one of them. Compared with depth-based representations, the GBR representation introduced in this paper, as well as the disparity and previous GBR, simplifies the depth by considering the predicted views. Moreover, the proposed GBR uses the concept of epipolar segment to keep the disparity unidimensional even for complex camera configurations. Finally, simplifications of the graph have been done to make the graph contain *just enough* information for constructing the *highest* quality predicted view (i.e., as the same quality as the rendering result by DIBR with uncompressed depth).

3. GBR REPRESENTATION

3.1. Graph Construction

Let us denote the constructed graph as $\mathcal{G}$, which has $2XY$ vertices corresponding to each pixel in $\mathcal{I}_1$ and $\mathcal{I}_2$. As mentioned before, pixels in $\mathcal{I}_2$ are grouped into a set of straight segments, pixels on each segment are supposed to have the same depth. Note that in this paper we select the straight segments horizontally (row by row, for example the y-th row of $\mathcal{I}_2$ has been divided into 3 straight segments in Fig.1), however it also works if the segments are chosen vertically (column by column), or in other order. Each segment has a connection that links itself (in $\mathcal{I}_2$) and its corresponding projection in $\mathcal{I}_1$. This edge thus describes the 3D geometry.

So far, the graph has been constructed with $2XY$ vertices and a number of connections (one connection for one segment). The color information is kept in the vertices, however only the first XY vertices corresponding to $\mathcal{I}_1$ (the reference view) need the color information. The connections of the graph represent the geometry information by providing the projection relation. Similarly to the disparity, the graph $\mathcal{G}$ simplifies the depth with respect to a given predicted view. We further consider the optimization of the graph (reducing the number of connections) while controlling the distortion of the rendered predicted view.

3.2. Graph Sparsification

According to the basic idea presented in section 3.1, pixels on each segment are represented with the same depth,

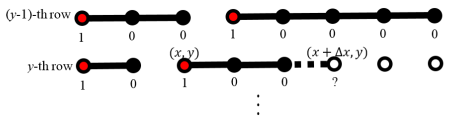


Fig. 2. Neighboring pixels grouping.

even though their initial depths are slightly different (within a range of Δz). However we can group pixels even with different depths, i.e., the segment depth varies within a range $[z-\Delta z, z+\Delta z]$ ($\Delta z \geq 0$). It is obvious that large Δz leads to an inaccurate geometry, which further leads to a low rendering quality. However rather than indirectly guaranteeing the rendering quality by setting Δz, we can select the segments according to the rendering quality.

As shown in Fig.2, each pixel in $\mathcal{I}_2$ has a label $c_2(x,y) = 1$ (the red pixels) or 0 (the dark pixels) which denotes whether pixel (x,y) is the beginning of a straight segment. Taking the segment beginning with pixel (x,y) as an example, pixel $(x+\Delta x, y)$ on the right is located on this segment if $c_2(x+\Delta x, y) = 0$, or another segment when $c_2(x+\Delta x, y) = 1$. Supposing that $c_2(x+\Delta x, y) = 0$, pixel $(x+\Delta x, y)$ is reconstructed with the same depth as the one of pixel (x,y) (since pixel $(x+\Delta x, y)$ is on the segment started by pixel (x,y)). Hence the rendering distortion of pixel $(x+\Delta x, y)$ can be given by

$$\mathcal{D}(x+\Delta x, y) = \left[I_2(x+\Delta x, y) - \hat{I}_2(x+\Delta x, y) \right]^2, \quad (3)$$

where the true color is $I_2(x+\Delta x, y)$ and the reconstructed one is $\hat{I}_2(x+\Delta x, y)$, which can be located by Eq.(1) with $(x+\Delta x, y)$ and the depth of pixel (x,y). Consequently, pixel $(x+\Delta x, y)$ can share the depth with pixel (x,y) if $\mathcal{D}(x+\Delta x, y)$ is small and it should start a new segment when $\mathcal{D}(x+\Delta x, y)$ is large. The computation of $c_2(x,y)$ can be given by

$$c_2(x+\Delta x, y) = \begin{cases} 1, & \text{if } \mathcal{D}(x+\Delta x, y) > \delta \\ 0, & \text{otherwise} \end{cases}. \quad (4)$$

In fact, Eq.(4) removes some connections which have limited contributions to the view construction quality.

3.3. Graph Coding

The connections of the graph can be represented by a huge binary matrix of size $XY \times XY$, which is a connectivity matrix between the $2XY$ pixels. A connection between two pixels is represented by 1 at associated position in the matrix. However, for each pixel in $\mathcal{I}_2$, all its possible projections (with varying depth) in $\mathcal{I}_1$ are located on an epipolar segment, as shown in Fig.3. In addition, the true projection (x', y') in $\mathcal{I}_1$ is between the *boundary* projections $(x'_{\min}, y'_{\min})$ and $(x'_{\max}, y'_{\max})$, where projection $(x'_{\min}, y'_{\min})$ is located by Eq.(1) using the minimum depth and projection $(x'_{\max}, y'_{\max})$ is related to the maximum depth [1]. Theoretically speaking, a real number is

[1] The minimum and maximum depths are the depth according to the minimum and maximum values given by the depth map.

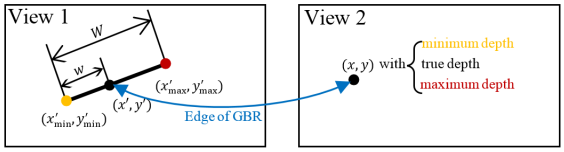
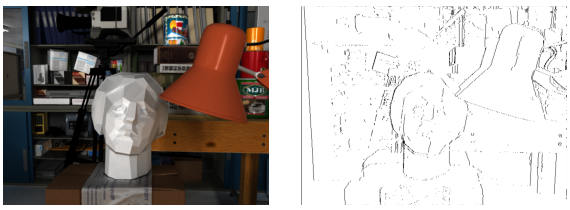


Fig. 3. w value is an index denoting the true projection position on the epipolar segment.



(a) View 2 (b) w map

Fig. 4. The predicted view and its associated w map.

enough to denote the projection position, e.g. the distance between $(x'_{\min}, y'_{\min})$ and (x', y'). Thus, a new quantity named w value denoting the projection position (x', y') is given by

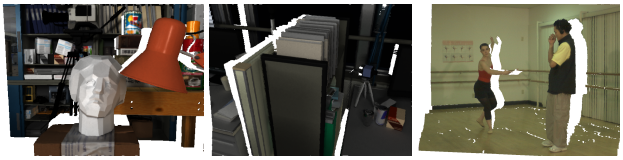
$$w = \begin{cases} -1, \text{ if pixel}(x,y) \text{ is new in } \mathcal{I}_2 \\ \text{round}\left(\frac{\sqrt{(x'_{\min}-x')^2+(y'_{\min}-y')^2}}{\sqrt{(x'_{\min}-x'_{\max})^2+(y'_{\min}-y'_{\max})^2}} W \right), \text{otherwise} \end{cases} \quad (5)$$

Note that $w = -1$ means that this segment (in $\mathcal{I}_2$) is new for reference view $\mathcal{I}_1$, such as the disoccluded or appearing pixels. By using the w values, the connectivity matrix can be replaced by a smaller binary matrix with size of $W \times XY$, where W depends on the quantization level of the epipolar segment. Since the w value measures the distance between $(x'_{\min}, y'_{\min})$ and (x', y'), it can be considered as *a disparity along the epipolar segment*. The accuracy of disparity on epipolar segments can reach to sub pixel with a large W, i.e., $W > \sqrt{(x'_{\min}-x'_{\max})^2 + (y'_{\min}-y'_{\max})^2}$.

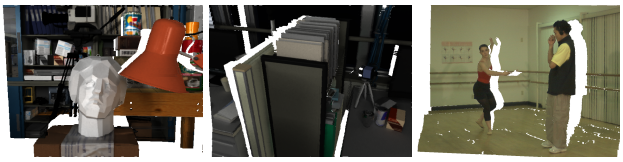
Fig.4.b shows the position of the pixels (in $\mathcal{I}_2$) with connections. Compared with the associated color image in Fig.4.a, we can see that the contours in the w map relate to the edges in the color image of $\mathcal{I}_2$. Based on this feature, we employ the arithmetic edge coding methods [12] to losslessly code the positions of these pixels. Accordingly, the differential pulse code modulation (DPCM) compression is used to code the w values along the edges. [13]

4. GBR VIEW RECONSTRUCTION

The graph described in section 3 is used to reconstruct the views. The reference view (i.e., view 1) is recovered directly by copying the color from the first XY vertices of the graph. The reconstruction of the predicted view (i.e., view 2) relies on the reconstruction of each segment. For a given segment, the following steps are applied: 1) The associated connection is recovered from the w value (locating the connection's endpoint on the epipolar segment by the w value); 2) The segment depth is estimated according to the connection by Eq.(1); 3) Then the first pixel of the segment is projected to



(a) Results by DIBR and depth maps compressed with HEVC (QP=10).



(b) Results by GBR.

Fig. 5. Reconstructed view 2. From left to right, dataset Tsukuba.1, Tsukuba.2 and MSR.Ballet. The blank regions are disocclusions with no color information copied from view 1.

the reference view by following the connection, meanwhile the last pixel of the segment is projected by Eq.(1) with the estimated depth; 4) The segment is interpolated with the color between the two projections in the reference view.

The *new* segments (related to the disoccluded or appearing pixels) is also known since their w values are -1. These segments are *holes* (no color information) in the reconstructed predicted view. The inpainting method proposed in [14] is used here to fill these holes.

5. EXPERIMENTS

In this section, the proposed GBR is comparatively assessed at the maximum rendering quality, i.e., the one obtained by DIBR and the original depth. Two datasets have been tested: 1) Tsukuba dataset [15]: 2 pairs of images have been used. For each pair of images, one is considered as reference view and the other is predicted view. 2) MSR dataset [16]: the Ballet dataset is tested. The image from camera 04 is the reference view and the image from camera 03 is the predicted view. The left of Tab.1 shows the camera translations between the reference and predicted views.

The proposed GBR is compared with two depth-based approaches, in which the depth maps are compressed with HEVC [17] and the method in [9]. In these baseline methods, the DIBR [18] method is applied to reconstruct the predicted views. Since the proposed construction of the GBR uses backward projection, in order to have a fair comparison, we also use DIBR with backward projection in the reference methods. The depth of the predicted view is first estimated by Eq.(1) with the decompressed depth of the reference view. Then the backward projection is applied with this estimated depth of the predicted view. For the proposed GBR, the depth of the predicted view is also estimated from the reference view. In this case, the proposed GBR, DIBR, HEVC and the method in [9] have the same input data and same pixel projection method.

W in Eq.(5) is selected as 255, which keeps the same quantization precision as the depth. The threshold of distortion in Eq.(3) varies with the tested datasets, which is around 650 (i.e., PSNR is around 20dB). The baseline approaches are tested with proper parameters to obtain highest reconstruction quality, e.g., HEVC with a QP parameter set to 0 or 10. Tab.1 lists the coding rate (in bitrate per pixel) obtained with different methods. The PSNR (with) measures the reconstruction quality of the whole predicted views (including the inpainting results), while the PSNR (no) only calculated on the pixels that are projected from the reference views (without considering the inpainting results). We can see from Tab.1 that the proposed GBR yields the highest reconstruction quality in terms of PSNR as well as the visual results (as shown in Fig.5) for a lower bit rate compared to DIBR with depth maps compressed with HEVC or method in [9]. The experiments show that the proposed GBR removes the redundant information in the depth, without a loss in rendering quality of the given predicted view.

Table 1. Coding rate (bits per pixel) of geometry and PSNR of the reconstructed predicted views by different methods. PSNR (with): PSNR considering the inpainting areas; PSNR (no): PSNR without considering the disoccluded areas.

Dataset (translation)	Methods		rate	PSNR (with)	PSNR (no)
Tsukuba.1 (forward translation, rotation)	DIBR		-	23.18	25.66
	HEVC	QP:0	0.358	23.19	25.66
		QP:10	0.186	23.22	25.57
	Method in [9]		0.381	23.18	25.66
	GBR		0.196	23.30	25.76
Tsukuba.2 (horizontal translation, rotation)	DIBR		-	28.24	31.67
	HEVC	QP:0	0.474	28.25	31.67
		QP:10	0.210	28.09	31.51
	Method in [9]		0.645	28.24	31.67
	GBR		0.233	28.22	31.93
MSR.Ballet (horizontal translation, rotation)	DIBR		-	28.34	31.32
	HEVC	QP:0	0.695	28.34	31.32
		QP:10	0.276	28.41	31.29
	Method in [9]		0.879	28.33	31.32
	GBR		0.347	28.57	31.44

6. CONCLUSION

In this paper, we have proposed an alternative to depth for multiview geometry representation. Contrary to the original GBR in [11], the proposed GBR can deal with multiview images with complex camera configurations. A distortion metric and disparity-based w value are proposed to simplify the graph. By these simple theories, the proposed GBR representation simplifies the depth of multiview images for reconstructing the given predicted views, i.e., the GBR costs less bitrate when obtaining the same high rendering quality.

Future works will focus on the full representation of both color and geometry, in which the connections of the graph should be used to form a better texture representation. Rate-distortion models should be investigated to consider both the rate cost and the distortion of the representation.

7. REFERENCES

[1] H. Y. Shum, S. B. Kang, and S. C. Chan, "Survey of image-based representations and compression techniques," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 11, pp. 1020–1037, 2003.

[2] K. Müller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 643–656, 2011.

[3] J. H. Park and H. W. Park, "Fast view interpolation of stereo images using image gradient and disparity triangulation," *Signal Processing: Image Communication*, vol. 18, no. 5, pp. 401–416, 2003.

[4] J. X. Chai, X. Tong, S. C. Chan, and H. Y. Shum, "Plenoptic sampling," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 2000, pp. 307–318.

[5] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Image Processing, IEEE International Conference on*, vol. 1. IEEE, 2007, pp. 201–204.

[6] E. Martinian, A. Behrens, J. Xin, and A. Vetro, "View synthesis for multiview video compression," in *Picture Coding Symposium*, vol. 37, 2006, pp. 38–39.

[7] K. R. Muller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F. H. Rhee, T. G., W. M., and W. T., "3D high-efficiency video coding for multi-view video and depth data," *Image Processing, IEEE Transactions on*, vol. 22, no. 9, pp. 3366–3378, 2013.

[8] H. Yuan, S. Kwong, J. Liu, and J. Sun, "A novel distortion model and Lagrangian multiplier for depth maps coding," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 24, no. 3, pp. 443–451, 2014.

[9] J. Gautier, O. Le Meur, and C. Guillemot, "Efficient depth map compression based on lossless edge coding and diffusion," in *Picture Coding Symposium (PCS), 2012*. IEEE, 2012, pp. 81–84.

[10] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H. 264/MPEG-4 AVC standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, 2011.

[11] T. Maugey, A. Ortega, and P. Frossard, "Graph-based representation for multiview image geometry," *Image Processing, IEEE Transactions on*, vol. 24, no. 5, pp. 1573–1586, 2015.

[12] I. Daribo, D. Florencio, and G. Cheung, "Arbitrarily shaped motion prediction for depth video compression using arithmetic edge coding," *Image Processing, IEEE Transactions on*, vol. 23, no. 11, pp. 4696–4708, 2014.

[13] T. Maugey, Y. H. Chao, A. Gadde, A. Ortega, and P. Frossard, "Luminance coding in graph-based representation of multiview images," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 130–134.

[14] J. Gautier, O. Le Meur, and C. Guillemot, "Depth-based image completion for view synthesis," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*. IEEE, 2011, pp. 1–4.

[15] M. Peris, A. Maki, S. Martull, Y. Ohkawa, and K. Fukui, "New Tsukuba Stereo Dataset." [Online]. Available: http://cvlab-home.blogspot.fr/2012/05/h2fecha-2581457116665894170-displaynone.html

[16] Interactive Visual Media group at Microsoft Research, "MSR 3D Video Dataset," 2014. [Online]. Available: http://research.microsoft.com/en-us/downloads/5e4675af-03f4-4b16-b3bc-a85c5bafb21d/

[17] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[18] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 93–104.